Assignment No. 4

4.1 Title

Consider a suitable text dataset. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors. Classify documents and evaluate precision, recall.

4.2 Problem Definition:

Remove stop words

4.3 Prerequisite:

Basic Concepts of ETL

4.4 Software Requirements:

Rapid Miner

4.5 Hardware Requirement:

PIV, 2GB RAM, 500 GB HDD, Lenovo A13-4089Model.

4.6 Learning Objectives:

We are going to learn how to tokenize and filter a document into its different words and then do words count for each word in a text document

4.7 Outcomes:

You are able to see a word list containing all the different words in your document and their occurrence count next to it in the "Total Occurrences" column

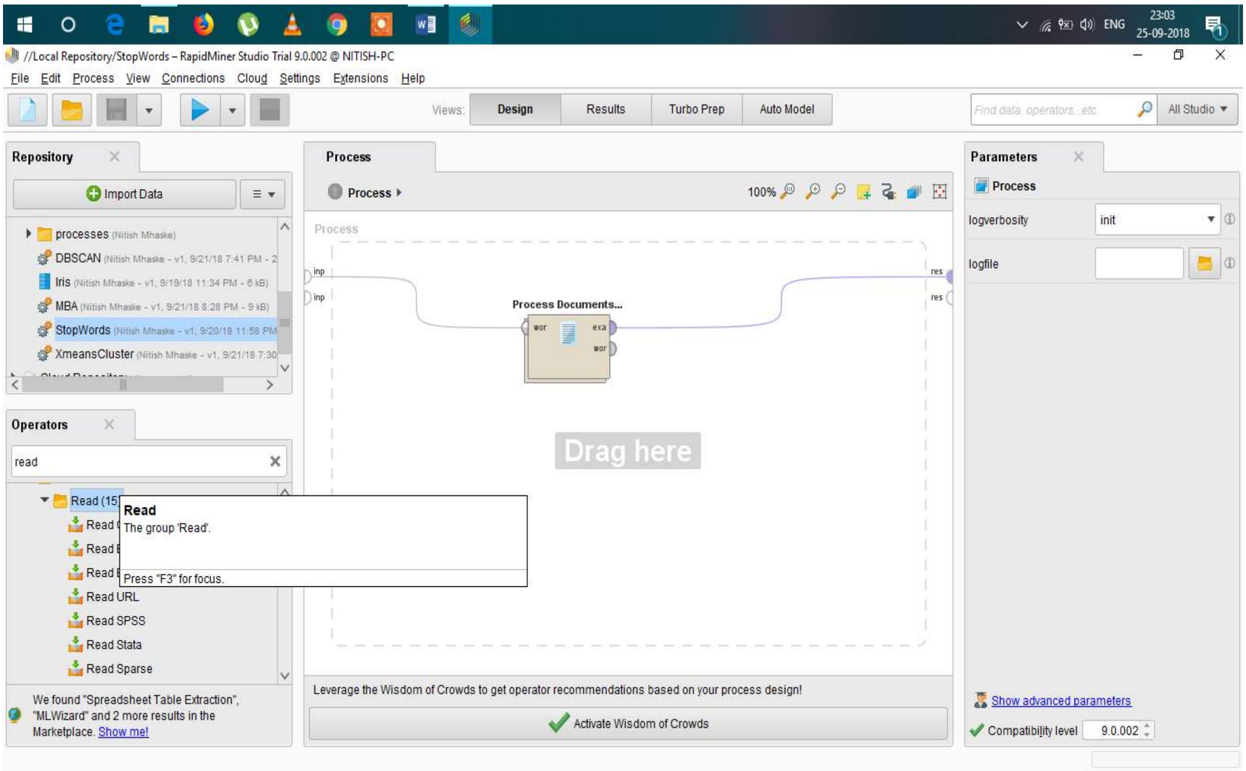
4.8 Theory Concepts:

4.8.1 Text Processing Tutorial with RapidMiner

In this, we are going to learn how to tokenize and filter a document into its different words and then do words count for each word in a text document.

Steps:

1. Open RapidMiner and click "New Process". On the left hand pane of your screen, there should be a tab that says "Operators"- this is where you can search and find all of the operators for RapidMiner and its extensions. By searching the Operators tab for "read".



//Local Repository/StopWords – RapidMiner Studio Trial 9.0.002 @ NITISH-PC

processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30

read

Read
The group 'Read'.
Press "F3" for focus.

Process Documents...

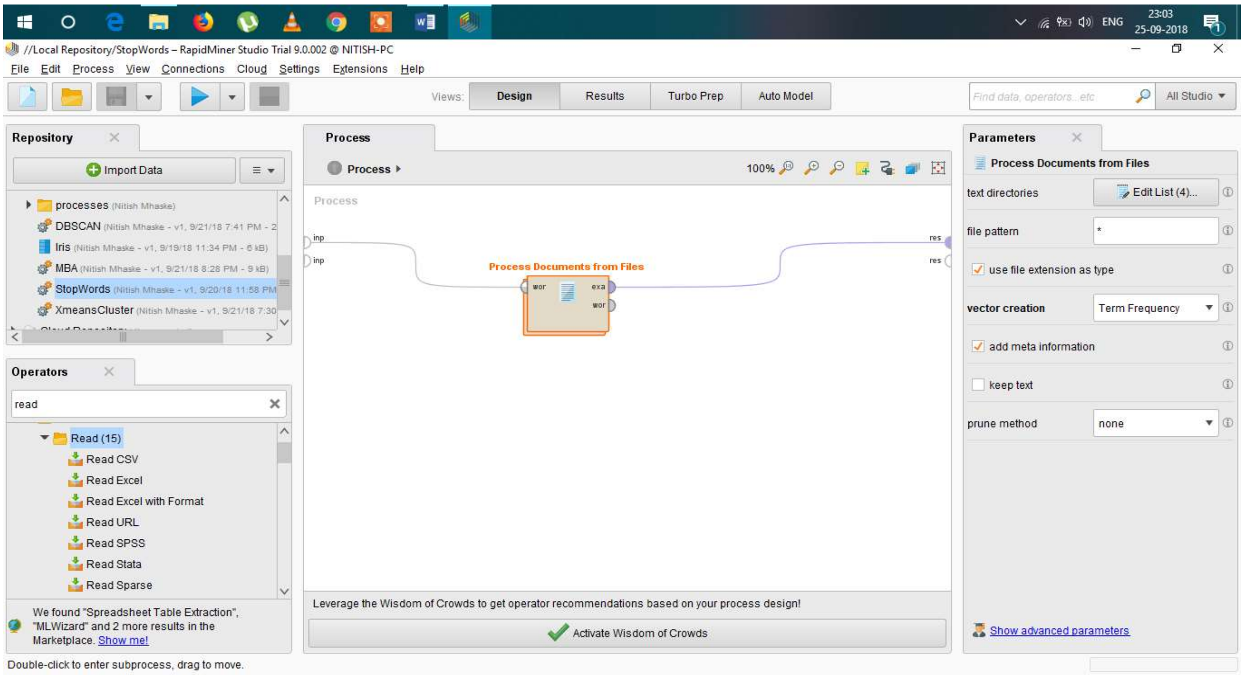
We found "Spreadsheet Table Extraction", "MLWizard" and 2 more results in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Process
logverbosity: init
logfile:
Compatibility level: 9.0.002

Figure 4.1: Searching the Operators tab for "read"

2. There are multiple read operators depending on which file you have, and most of them work the same way. If you scroll down, there is a "Read Documents" operator. Select this operator and enter it into your Main Process window by dragging it. When you select the Read Documents operator in the Main Process window, you should see a file up loader in the right-hand pane.



23:03
25-09-2018
ENG
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Operators
Process
100%
inp
res
Parameters
Activate Wisdom of Crowds
Show advanced parameters

//Local Repository/StopWords – RapidMiner Studio Trial 9.0.002 @ NITISH-PC

processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30

read

Read (15)
Read CSV
Read Excel
Read Excel with Format
Read URL
Read SPSS
Read Stata
Read Sparse

Process Documents from Files

We found "Spreadsheet Table Extraction", "MLWizard" and 2 more results in the Marketplace. Show me!

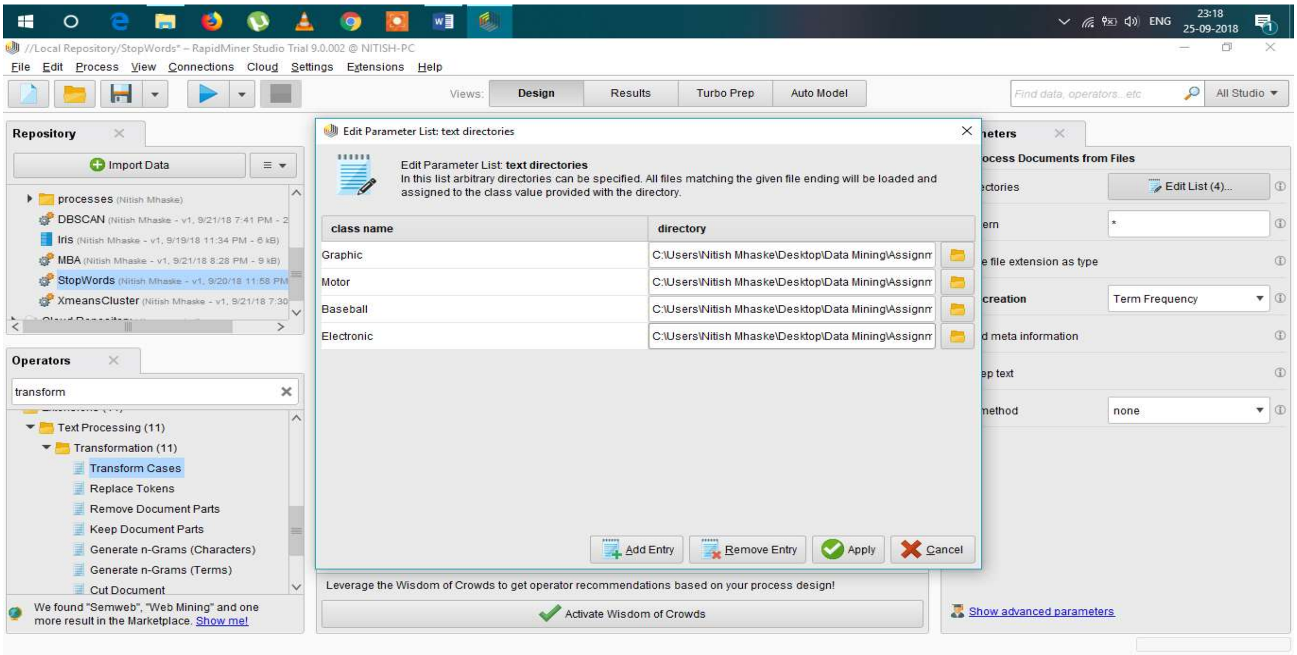
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Process Documents from Files
text directories: Edit List (4)...
file pattern: *
☑ use file extension as type
vector creation: Term Frequency
☑ add meta information
□ keep text
prune method: none

Double-click to enter subprocess, drag to move.

Figure 4.2: Drag and Drop “Process Documents" operator

3. Select the text file you want to use.



Edit Parameter List: text directories

Edit Parameter List: **text directories**
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.

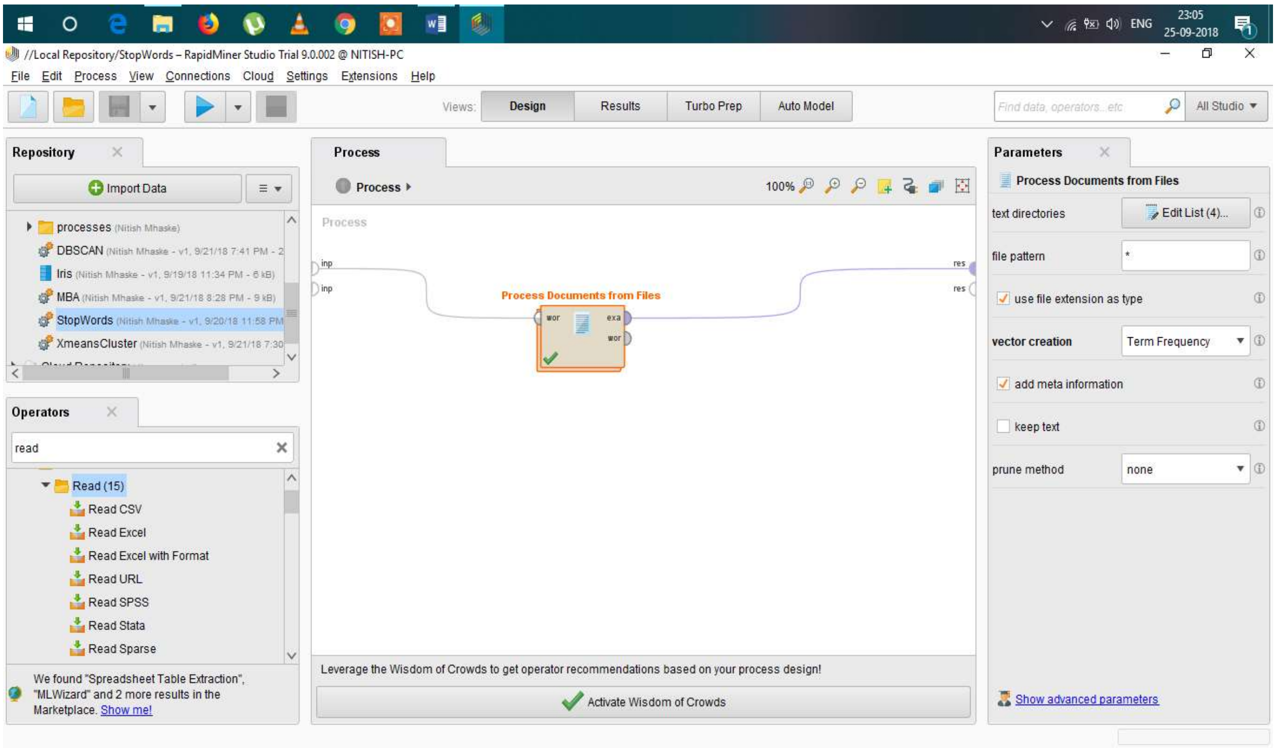
class name	directory
Graphic	C:\Users\Nitish Mhaske\Desktop\Data Mining\Assignm
Motor	C:\Users\Nitish Mhaske\Desktop\Data Mining\Assignm
Baseball	C:\Users\Nitish Mhaske\Desktop\Data Mining\Assignm
Electronic	C:\Users\Nitish Mhaske\Desktop\Data Mining\Assignm

We found "Semweb", "Web Mining" and one more result in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 4.3: Select the text file you want to use

4. After you have chosen your file, make sure that the output port on the Read Documents operator is connected to the "res" node in your Main Process. Click the "play" button to check that your file has been received correctly. Switch to the results perspective by clicking the icon that looks like a display chart above the "Process" tab at the top of the Main Process pane. Click the "Document (Read Document)" tab. Your output text should look something like this depending on the file you have chosen to process:


//Local Repository/StopWords – RapidMiner Studio Trial 9.0.002 @ NITISH-PC
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
processes
DBSCAN
Iris
MBA
StopWords
XmeansCluster
Operators
read
Read (15)
Read CSV
Read Excel
Read Excel with Format
Read URL
Read SPSS
Read Stata
Read Sparse
Process
100%
inp
inp
res
res
Process Documents from Files
wor
exa
wor
Activate Wisdom of Crowds
Show advanced parameters
23:05
25-09-2018

We found "Spreadsheet Table Extraction", "MLWizard" and 2 more results in the Marketplace. Show me!

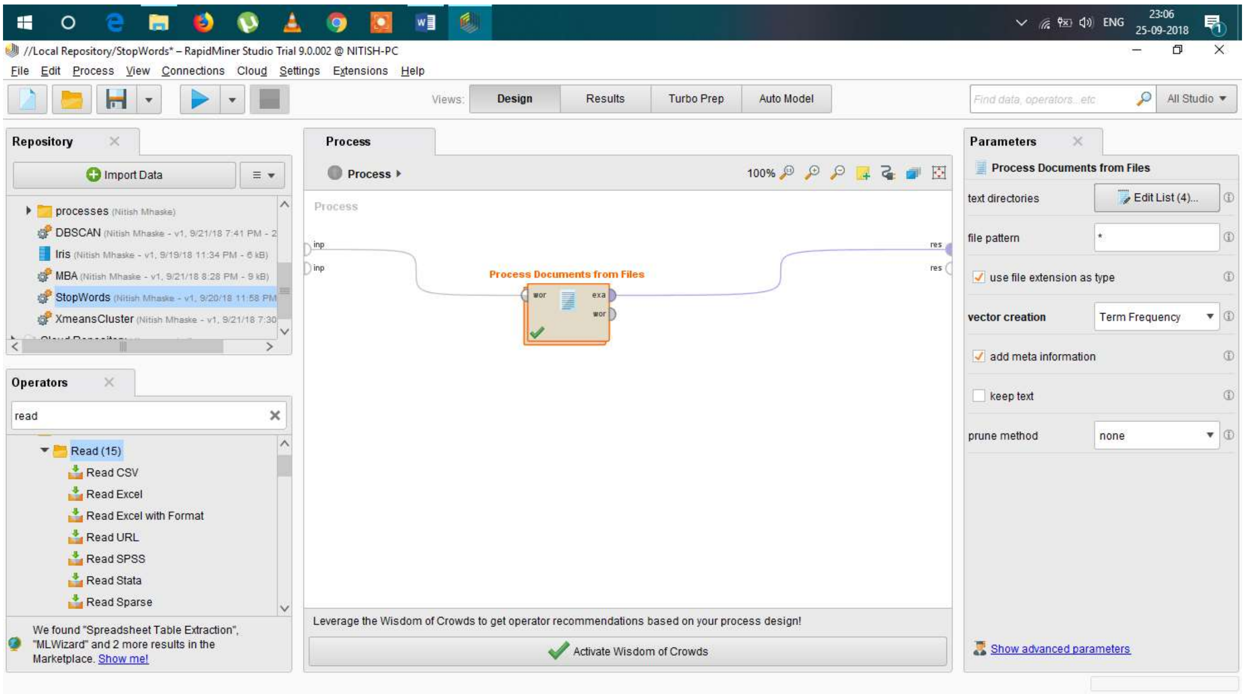
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Parameters — Process Documents from Files

text directories	Edit List (4)...
file pattern	*
☑ use file extension as type	
vector creation	Term Frequency
☑ add meta information	
☐ keep text	
prune method	none

Figure 4.4: Run the Process

5. Now we will move on to processing the document to get a list of its different words and their individual count. Search the Operators list for "Process Documents". Drag this operator the same way as you did for the "Read Documents" operator into the main panel.



//Local Repository/StopWords* – RapidMiner Studio Trial 9.0.002 @ NITISH-PC

Parameters: Process Documents from Files

text directories: Edit List (4)...

file pattern: *

☑ use file extension as type

vector creation: Term Frequency

☑ add meta information

□ keep text

prune method: none

Operators: read

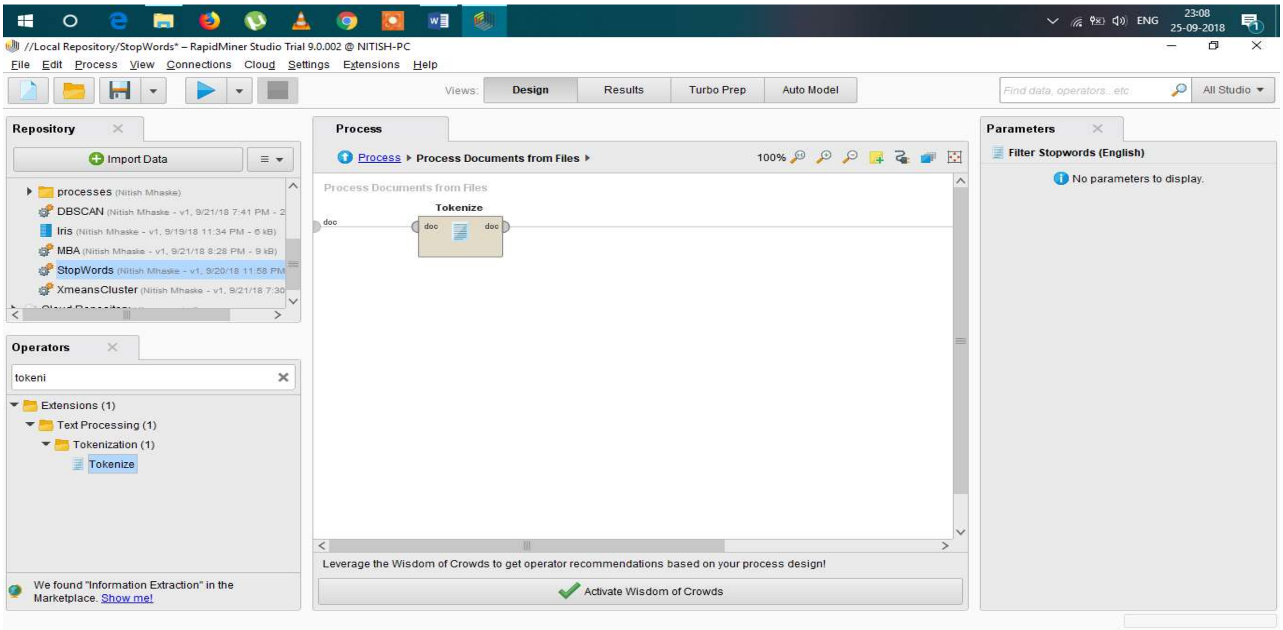
We found "Spreadsheet Table Extraction", "MLWizard" and 2 more results in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30
Process
Process ▸
100%
inp
inp
res
res
Process Documents from Files
wor
exa
wor
Read (15)
Read CSV
Read Excel
Read Excel with Format
Read URL
Read SPSS
Read Stata
Read Sparse
Activate Wisdom of Crowds
Show advanced parameters
ENG
23:06
25-09-2018

Figure 4.5: Search the Operators list for "Process Documents"

6. Double click the Process Documents operator to get inside the operator. This is where we will link operators together to take the entire text document and split it down into its word components. This consists of several operators that can be chosen by going into the Operator pane and looking at the Text Processing folder. You should see several more folders such as "Tokenization", "Extraction", "Filtering", "Stemming", "Transformation", and "Utility".



//Local Repository/StopWords* – RapidMiner Studio Trial 9.0.002 @ NITISH-PC

Parameters: Filter Stopwords (English)

No parameters to display.

Operators: tokeni

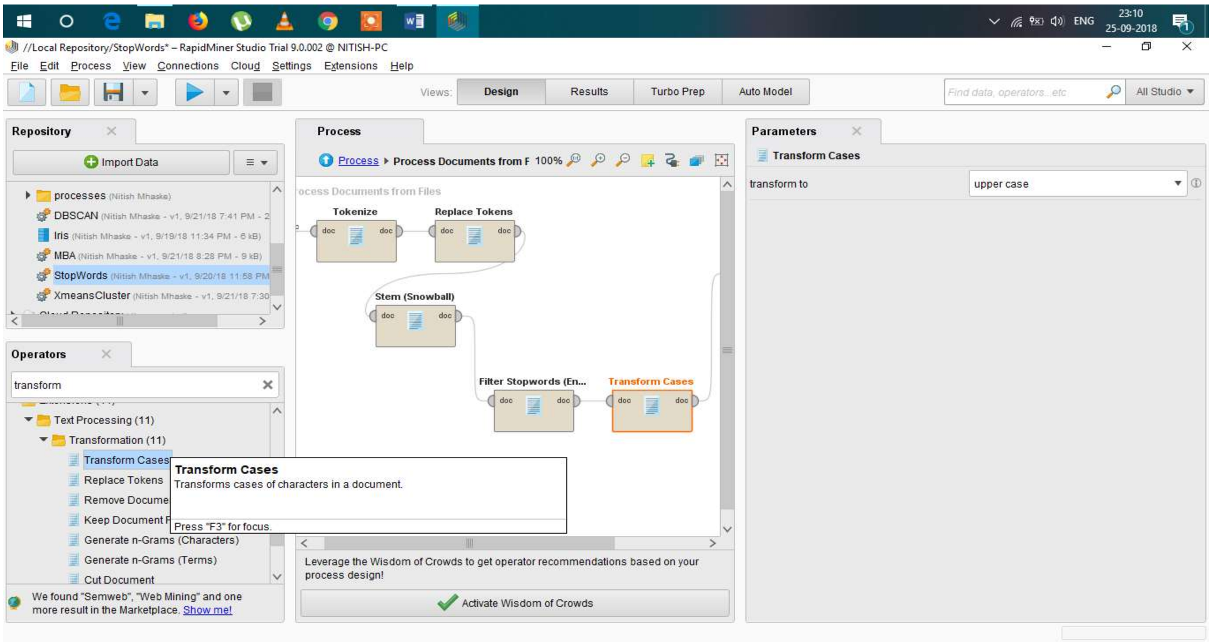
We found "Information Extraction" in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30
Process
Process ▸ Process Documents from Files ▸
100%
Process Documents from Files
Tokenize
doc
doc
doc
Extensions (1)
Text Processing (1)
Tokenization (1)
Tokenize
Activate Wisdom of Crowds
ENG
23:08
25-09-2018

Figure 4.6 Search for the "Tokenize" operator

7. Connect the "doc" node of the process to the "doc" input node of the operator if it has not automatically connected already. Now we are ready to filter the bag of words. In "Filtering" folder under the "Text Processing" operator folder, you can see the various filtering methods that you can apply to your process.



//Local Repository/StopWords* – RapidMiner Studio Trial 9.0.002 @ NITISH-PC
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30
Operators
transform
Text Processing (11)
Transformation (11)
Transform Cases
Replace Tokens
Remove Docume
Keep Document P
Generate n-Grams (Characters)
Generate n-Grams (Terms)
Cut Document
Process
Process ▸ Process Documents from F 100%
Process Documents from Files
Tokenize
Replace Tokens
Stem (Snowball)
Filter Stopwords (En...
Transform Cases
doc
Parameters
Transform Cases
23:10
25-09-2018
ENG

Transform Cases
Transforms cases of characters in a document.

Press "F3" for focus.

transform to: upper case

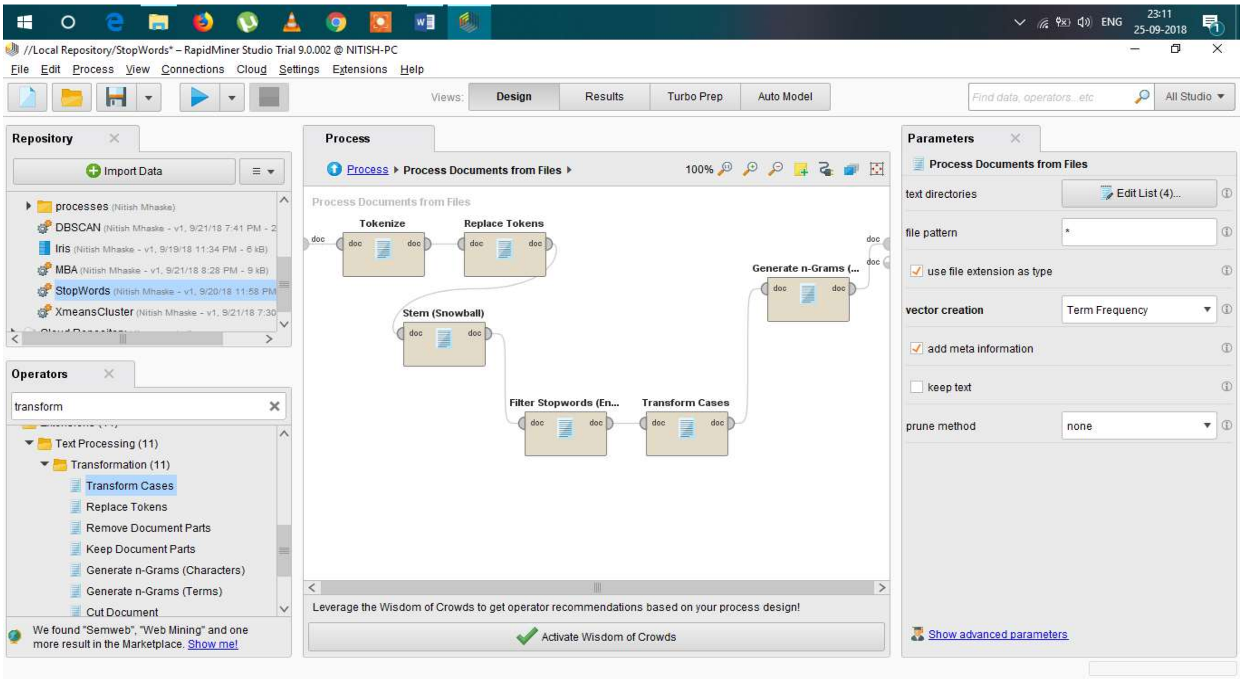
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

We found "Semweb", "Web Mining" and one more result in the Marketplace. Show me!

Figure 4.7 Select the operator "Transform Cases" and drag it into the process.

8. After I filtered the bag of words by stopwords and length, I want to transform all of my words to lowercase since the same word would be counted differently if it was in uppercase vs. lowercase. Select the operator "Transform Cases" and drag it into the process.



//Local Repository/StopWords* – RapidMiner Studio Trial 9.0.002 @ NITISH-PC
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
processes (Nitish Mhaske)
DBSCAN (Nitish Mhaske - v1, 9/21/18 7:41 PM - 2
Iris (Nitish Mhaske - v1, 9/19/18 11:34 PM - 6 kB)
MBA (Nitish Mhaske - v1, 9/21/18 8:28 PM - 9 kB)
StopWords (Nitish Mhaske - v1, 9/20/18 11:58 PM
XmeansCluster (Nitish Mhaske - v1, 9/21/18 7:30
Operators
transform
Text Processing (11)
Transformation (11)
Transform Cases
Replace Tokens
Remove Document Parts
Keep Document Parts
Generate n-Grams (Characters)
Generate n-Grams (Terms)
Cut Document
Process
Process ▸ Process Documents from Files ▸
100%
Process Documents from Files
Tokenize
Replace Tokens
Stem (Snowball)
Filter Stopwords (En...
Transform Cases
Generate n-Grams (...
doc
Parameters
Process Documents from Files
23:11
25-09-2018
ENG

text directories: Edit List (4)...

file pattern: *

☑ use file extension as type

vector creation: Term Frequency

☑ add meta information

□ keep text

prune method: none

Show advanced parameters

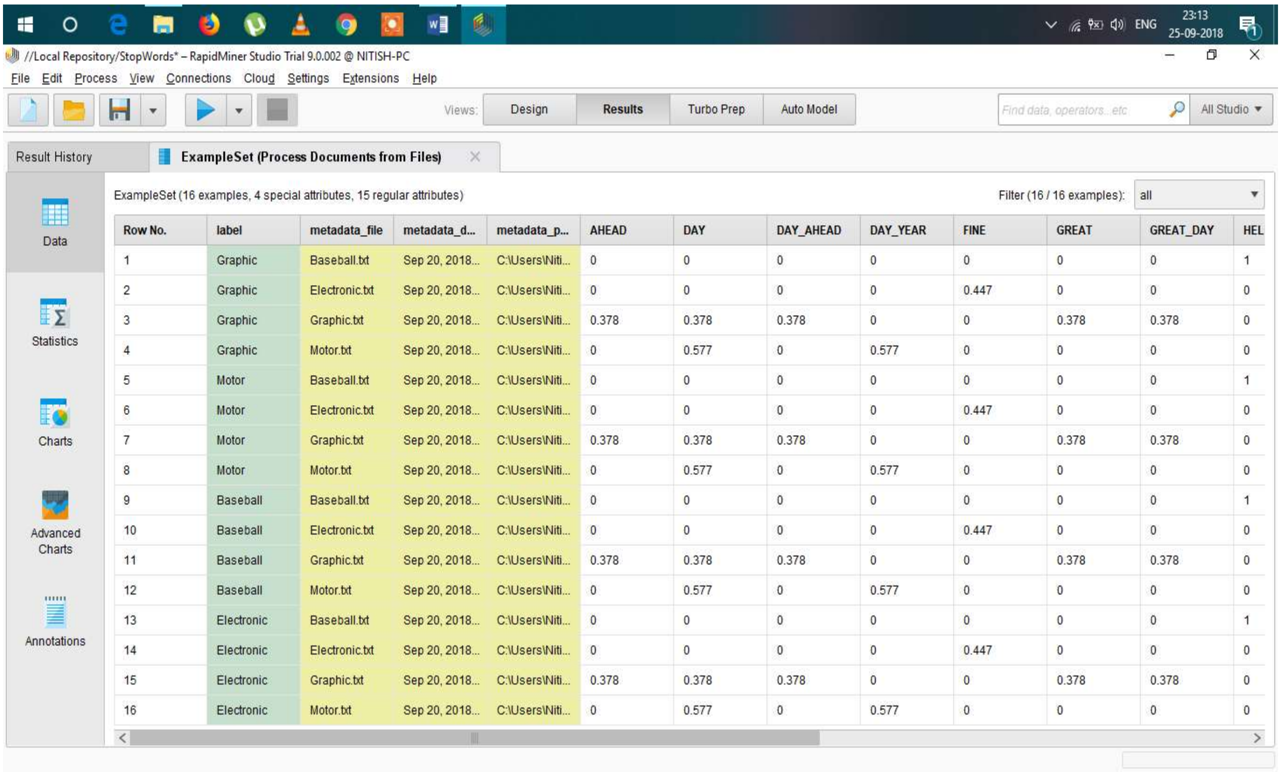
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

We found "Semweb", "Web Mining" and one more result in the Marketplace. Show me!

Figure 4.8 Checks all nodes connections and clicks the "Play" button to run process

9. Now that I have the sufficient operators in my process for this example, I check all of my node connections and click the "Play" button to run my process. If all goes well, your output should look like this in the results view:



23:13
25-09-2018
ENG
//Local Repository/StopWords* – RapidMiner Studio Trial 9.0.002 @ NITISH-PC
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Process Documents from Files)
Data
Statistics
Charts
Advanced Charts
Annotations

ExampleSet (16 examples, 4 special attributes, 15 regular attributes)

Filter (16 / 16 examples): all

Row No.	label	metadata_file	metadata_d...	metadata_p...	AHEAD	DAY	DAY_AHEAD	DAY_YEAR	FINE	GREAT	GREAT_DAY	HEL
1	Graphic	Baseball.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0	0	0	1
2	Graphic	Electronic.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0.447	0	0	0
3	Graphic	Graphic.txt	Sep 20, 2018...	C:\Users\Niti...	0.378	0.378	0.378	0	0	0.378	0.378	0
4	Graphic	Motor.txt	Sep 20, 2018...	C:\Users\Niti...	0	0.577	0	0.577	0	0	0	0
5	Motor	Baseball.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0	0	0	1
6	Motor	Electronic.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0.447	0	0	0
7	Motor	Graphic.txt	Sep 20, 2018...	C:\Users\Niti...	0.378	0.378	0.378	0	0	0.378	0.378	0
8	Motor	Motor.txt	Sep 20, 2018...	C:\Users\Niti...	0	0.577	0	0.577	0	0	0	0
9	Baseball	Baseball.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0	0	0	1
10	Baseball	Electronic.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0.447	0	0	0
11	Baseball	Graphic.txt	Sep 20, 2018...	C:\Users\Niti...	0.378	0.378	0.378	0	0	0.378	0.378	0
12	Baseball	Motor.txt	Sep 20, 2018...	C:\Users\Niti...	0	0.577	0	0.577	0	0	0	0
13	Electronic	Baseball.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0	0	0	1
14	Electronic	Electronic.txt	Sep 20, 2018...	C:\Users\Niti...	0	0	0	0	0.447	0	0	0
15	Electronic	Graphic.txt	Sep 20, 2018...	C:\Users\Niti...	0.378	0.378	0.378	0	0	0.378	0.378	0
16	Electronic	Motor.txt	Sep 20, 2018...	C:\Users\Niti...	0	0.577	0	0.577	0	0	0	0

Figure 4.9 Output should look like this in the results view

4.9 Conclusion

We are now able to see a word list containing all the different words in your document and their occurrence count next to it in the "Total Occurrences" column. If you do not get this output, make sure that all of your nodes are connected correctly and also to the right *type*. Some errors are because your output at one node does not match the type expected at the input of the next node of an operator

References:- https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf